Project 09 Writeup

Data Platform

At this stage of the project, the first step is to check out the data set and begin compiling statistics on the imported data. Due to the relational nature of the data, the first step was to generate a database implementing these relations. I chose to write my preprocessing script in Ruby with ActiveRecord to load data into a MySQL database. This process was extraordinarily slow (especially with SQLite), and was run overnight.

Statistics

Using the database, it was easy to find the number of each object (1391 albums, 2487 artists, 479 genres, 7295 tracks) and finding the average rating by userid. and the average rating of each track by genre by user. These allowed me to find basic numbers on each user, which could be used as a crude prediction algorithm.

Clustering

Users could be clustered given their average ratings on genres/artists/albums. This would require extra work in preprocessing the data.

Another interesting statistic might be to cluster on ratings of tracks from the same genre.

Classification

The most useful classifier I can think of for this data is an ANN. Given the input set of data and having all tracks somewhat well represented with at least ten ratings gives a lot of input data for an ANN. Additionally, it would function well to predict scores of users as the project requires.

Conclusions

I would use a ANN classifier for this task as a first attempt. Perhaps this could then be made into an ensemble method combining an ANN with other classification strategies such as KNN or doing some sort of probabilistic approach. In the winner's project slideshows, they spoke of Restricted Boltzmann Machines which are a form of recurrent neural network. They are quite a bit more complex and would require a lot of implementation to figure out.